The performance of deep learning tools, trained on the same dataset, for auto-segmentation of challenging organs-at-risk in the thoracic region

Sevgi Emin[1], Elia Rossi[2], Mattias Hedman[2,3], Marcela Giovenco[2], Daniel Alm[2], Fernanda Villegas[1,3], Eva Onjukka[1,3]

[1]Nuclear Medicine and Medical Physics, Karolinska University Hospital, Stockholm, Sweden. [2]Department of Radiation Oncology, Karolinska University Hospital, Stockholm, Sweden. [3]Department of Oncology-Pathology, Karolinska Institute, Stockholm, Sweden

Purpose/Objective:

The critical factors for training successful AI-based models for auto-segmentation remain unclear, preventing a meaningful comparison of commercially available tools. The purpose of this study is to compare the performance of three AI-based auto-segmentation models for the thoracic region, where the same dataset has been available for training. Organs-at-risk (OAR) auto-segmentation of the thoracic region is of particular interest as it provides a range of challenging anatomical structures, potentially highlighting any strengths and weaknesses between the three model training approaches.

Material/Methods:

For 250 lung cancer patients, a structure set was delineated and reviewed by Radiation Oncology experts. The patients were divided into two datasets: 200 for training and 50 for testing. Three companies participated in the study, each having access to the training dataset for a limited time to develop a model according to their method of choice. The three models were tested on the blind test dataset by the authors. The quantitative analysis employed the metrics: volumetric Dice Similarity Coefficient (DSC), mean Hausdorff distance (mHD) and surface DSC (sDSC). The latter with a tolerance of 1mm. Inter-observer variability in manual segmentation was estimated by three independent expert delineations for a subset of 5 test patients following the same guidelines [1,2]. The mean values from the observers were used to normalize the auto-segmentation results as described by Yang et al [3], where a score of 100 represents a perfect value, whilst a score of 50 is equivalent to the inter-observer variability.

Results:

Good performance was observed for all three models for most structures as seen in Figure 1. Based on the low DSC and sDSC and high mHD, the most challenging structures were the brachial plexus, pulmonary vein and inferior vena cava. Structures such as the aorta, thoracic wall and pulmonary artery showed a high DSC, sDSC and low mHD with low confidence intervals for all metrics. Larger confidence intervals were observed for structures in areas with low contrast, partially due to more outliers.

The results normalized to inter-expert variability (Figure 2) suggest a similar performance between the models for all metrics. All mean values were above 50, indicating a smaller deviation from the reference delineation compared to the inter-observer variability.

The performance of deep learning tools, trained on the same dataset, for auto-segmentation of challenging organs-at-risk in the thoracic region

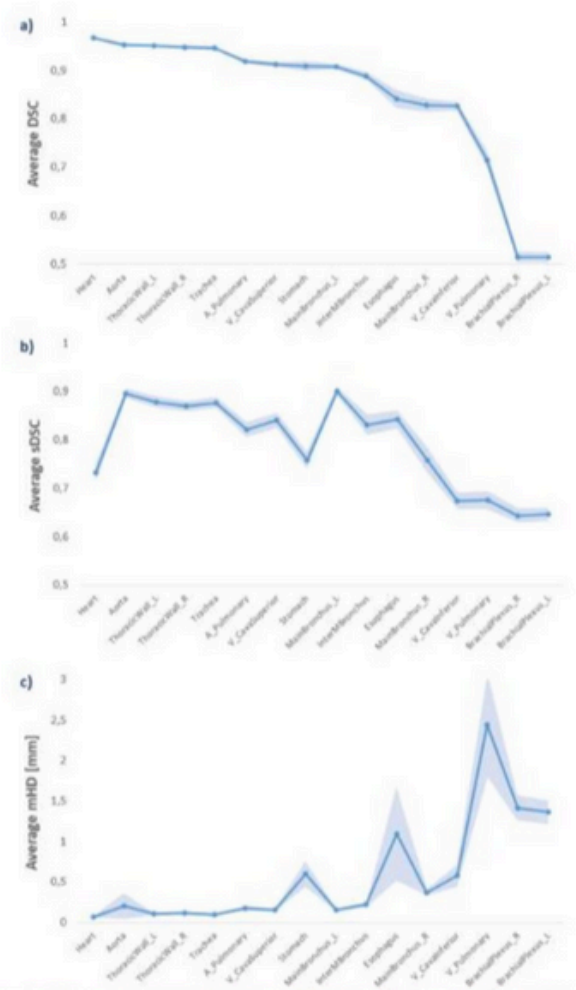

Figure 1: The average a) DSC, b) sDSC, and c) mHD with the respective ± 95% confidence interval for all patients and models. The order in the x-axis follows the order from figure a).

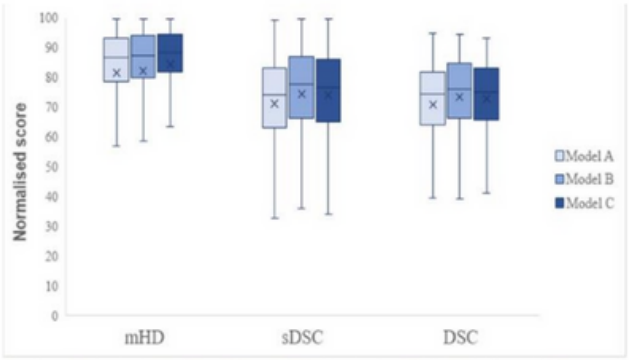

Figure 2: Normalised mHD, sDSC and DSC scores for all contours per model. A score of 100 represents a perfect value, whilst a score of 50 is equivalent to the inter-observer variability.

Conclusion:
Overall, the three auto-segmentation models showed similar performance, and all were challenged by both branching organs and organs in areas with lower soft tissue contrast. The auto-segmentation tools exhibited higher consistency compared to inter-observer variations.

Keywords: thorax, comparison, deep-learning